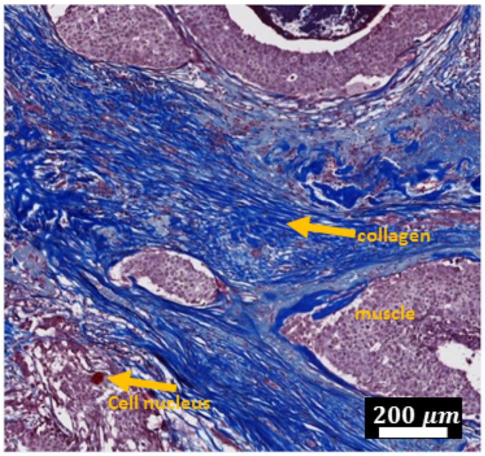
Classification of Infrared Spectroscopic Images

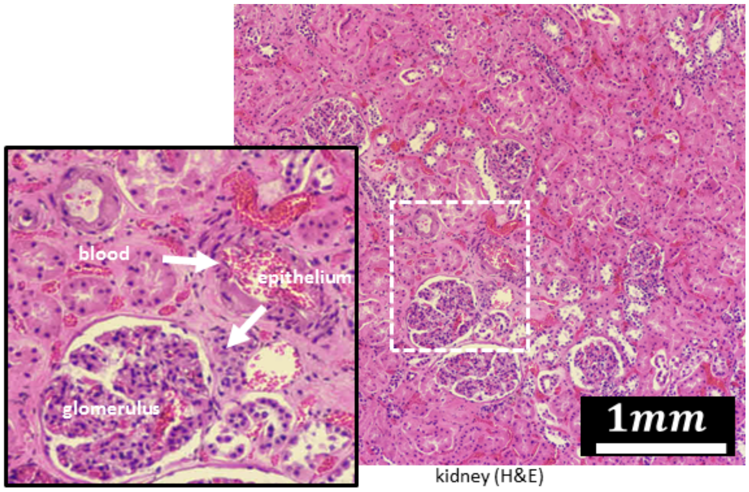
1. High level description of project.

Histopathology is the gold standard for cancer diagnosis and determining initial directions for treatment. The high level of morphological detail present in stained biopsies enables pathologists to determine the presence of cancer.

Histological stains are used for microscopic visual examination of the tissue structures:

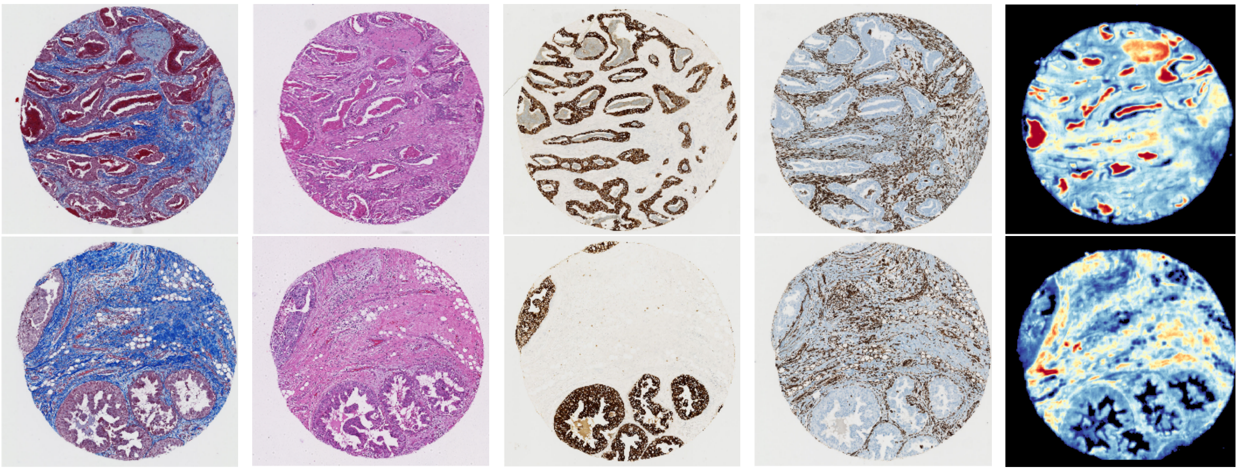


Human breast stained with Masson's trichrome



kidney (H&E)

There are alternative approaches such as Fourier transform infrared spectroscopy (FTIR). The absorption spectra provide molecular fingerprints for each pixel, which translates to key cellular biomolecules, such as proteins, lipids, DNA, collagen, glycogen, and carbohydrates. Each pixel contains the spectrum generated by a Fourier Transform.



(a)Masson's trichrome

(b)H & E

(c)Cytokeratin

(d)Vimentin

(e)Mid-infrared (1 band)

Figure: Chemically stained (a-d), (e) Colormapped mid-infrared images of the corresponding two cores are shown, where color indicates the magnitude of the absorbance spectrum in arbitrary units at 1650cm^{-1}

2. **What question or problem are you trying to solve?**
Standard histopathology steps consist of biopsy collection, tissue preparation and sectioning, the application of chemical stains, and analysis by an expert pathologists. This process is done manually and has some problems:
    - time-consuming and susceptible to human error
    - Chemical stains are non-quantitative
    - heavily relied upon in clinical assessment

   The goal is to identify six major cellular or acellular constituents of tissue (namely blood, collagen, epithelium, necrosis, myofibroblasts and adipocytes) using spectral information from FTIR.

3. **How will you present your work?**
I would like to use interactive visualization if I have enough time, otherwise I stay with PowerPoint presentation.

4. **What are your data sources?**
The dataset used in this study consisted of 540 breast tissue cores (with 1mm diameter) from different patients.

   Feature: The dataset comprised of 680 IR- spectra (680 bands). 680 features for each pixel.

   Target: Histological sections were examined by experienced pathologists to identify cell types within the tissues (blood, collagen, epithelium, necrosis, myofibroblasts and adipocytes).

Classes	Data
Blood	10,942
Collagen	1,431,663
Epithelium	451,716
Necrosis	163,094
myofibroblasts	592,344
adipocytes	55,000

5. **What's your next step towards making this your project?**
    - Preparing data (I think I need to balance data for different classes)
    - Apply PCA for dimension reduction